认知任务的信度：现状与展望

朱芃芃[#]
(南京师范大学心理学院，南京，210024; 江苏省高校哲学社会科学实验室——南京师范大学青少年教育与智能支持实验室，南京，210024)
刘铮[#]
(香港中文大学(深圳)人文社科学院，深圳，518172)
康春花[*]
(浙江省儿童青少年心理健康与危机干预智能实验室，金华，321004)
胡传鹏[*1]
(南京师范大学心理学院，南京，210024; 江苏省高校哲学社会科学实验室——南京师范大学青少年教育与智能支持实验室，南京，210024)

摘　要 认知任务作为认知心理学的核心方法，近年来被开始用于测量个体的认知差异，但认知任务的信度估计方法及信度研究现状仍然较少被关注。本文从传统的信度概念出发，在指出研究者对认知任务信度理解和评估误区的基础上，概括了认知任务信度评估面临的两大挑战。通过梳理和总结认知任务信度评估的计算方法及指标选取，进而详细探讨了提升认知任务信度的三大策略，即增加被试间差异；降低误差分数分数方差；改良计算与统计方法。本文从认知任务信度概念和应用的混淆误区，适宜的评估方法和指标选取，面临的两大挑战和信度提升策略，系统地对认知任务信度进行了梳理和总结，以期为研究者提供合理适宜的方法选择和应用路径。

关键词　认知任务，信度悖论，信度，个体差异

英文摘要

Cognitive tasks are central to experimental psychology and cognitive neuroscience, offering valuable insights into human cognition, decision-making, and behavior. These tasks are widely utilized in research to investigate cognitive mechanisms and are also integrated into clinical assessments to examine cognitive dysfunction in psychiatric and neurological disorders. However, one often overlooked but critical issue is the reliability of these tasks in measuring individual differences.

Reliability, a fundamental psychometric property, refers to the consistency of a measure across different testing occasions, repetitions, and observers. In traditional psychometric assessments, reliability is a crucial criterion for evaluating the validity of a measurement tool. However, in cognitive psychology, the reliability of cognitive tasks has not received sufficient attention, partly due to the historical divide between experimental psychology and

[1] 通讯作者：胡传鹏，email: hcp4715@hotmail.com; 康春花，email: akang@zjnu.cn

psychometric research. As cognitive neuroscience increasingly integrates these tasks to explore brain-behavior relationships, understanding and improving task reliability is essential for advancing both theoretical and applied research.

This paper begins by revisiting the traditional concept of reliability and identifying common misconceptions in understanding and assessing the reliability of cognitive tasks. We then summarize two major challenges in evaluating cognitive task reliability: (1) the inherent variability in cognitive performance, which complicates reliability estimation, and (2) the limited applicability of conventional psychometric methods in cognitive task assessments. To address these challenges, we review existing computational methods and reliability indices commonly used in cognitive task reliability studies, including split-half reliability and intraclass correlation coefficients (ICC).

Traditional internal consistency measures, such as Cronbach's alpha, are often inadequate for cognitive tasks due to their inability to account for trial-wise variability. Instead, split-half reliability is preferred because it assesses internal consistency by dividing the test into two halves and evaluating their correlation. Pronk et al. (2022) systematically evaluated four commonly used split-half methods and found that permutation-based split-half reliability was the most robust approach. This method involves repeated random sampling without replacement to divide trials into two halves and computing the Spearman-Brown-corrected Pearson correlation for each split. By generating a distribution of reliability coefficients, this approach mitigates sampling errors from a single split while accounting for potential time effects, task design influences, and nonlinear transformations in cognitive experiments.

Another preferred method for evaluating test-retest reliability in cognitive tasks is the Intraclass Correlation Coefficient (ICC), which assesses both correlation and agreement in repeated measurements. There are various ICC models, with McGraw and Wong (1996) outlining 10 different variants, making it challenging to select the appropriate one. In cognitive task reliability studies, the most recommended models are ICC(2,1)and ICC(3,1) (Parsons et al., 2019). ICC(2,1), derived from a two-way random-effects model, is suitable when generalizing results to a broader population, while ICC(3,1), based on a two-way mixed-effects model, applies when the results pertain only to the specific sample tested (Koo & Li, 2016). To ensure robustness, researchers often report both models for comparison. Reliability metrics are generally interpreted as follows: ICC < 0.5 indicates poor reliability, 0.5–0.75 represents moderate reliability, 0.75–0.9 suggests good reliability, and ICC > 0.9 signifies excellent reliability (Cicchetti & Sparrow,

1981; Kupper & Hafner, 1989). However, these thresholds should be considered within the context of the measurement tool and sample characteristics.

We next review empirical findings on the reliability of widely used cognitive tasks. Reliability estimates vary significantly across studies, depending on task design and response metrics. Finally, we discuss three key strategies to improve the reliability of cognitive tasks: (1) increasing inter-individual variability to enhance measurement sensitivity, (2) reducing error variance to improve measurement precision, and (3) refining computational and statistical methods for reliability estimation.

In conclusion, ensuring the reliability of cognitive tasks is crucial for their effective application in both research and clinical settings. Future research should continue to refine reliability assessment methods, explore alternative computational modeling approaches, and develop innovative experimental paradigms that balance ecological validity with psychometric rigor. By systematically analyzing the concept of cognitive task reliability, common evaluation methods, key challenges, and improvement strategies, this paper aims to provide researchers with appropriate methodological choices and practical guidelines for enhancing the reliability of cognitive task measurements.

1 引言

认知任务在实验和认知心理学中扮演着揭示认知机制的关键角色。近年来，随着计算机算力的飞速提升与数据可得性的增加，计算建模在认知科学领域获得了广泛关注(Kriegeskorte & Douglas, 2018)，增强了认知过程量化的精度。特别是在精神病学研究中，计算建模结合认知任务加深了人们对认知异常和个体差异的理解(区健新 等, 2020; Huys, 2015; Montague et al., 2012)，推动了个体化诊疗策略的发展(Geng et al., 2022; Huys et al., 2021)。

然而，认知任务能否可靠地测量个体的认知差异，即认知信度，却一直未受到足够重视(如，Dang et al., 2020; Parsons et al., 2019; Yarkoni & Braver, 2010)。这一问题部分源于科学心理学历史上实验心理学与心理测量的分离(Cronbach, 1957)：实验心理学强调群体水平的实验效应稳定性，而心理测量学则致力于开发衡量个体差异的工具和方法(Vasey et al., 2003)。这一分离意味着实验心理学研究者不太关注认知任务在测量个体差异上的表现。此外，大多数认知任务在群体水平上表现出较好的可重复性，但群体信度与衡量个体差异的信度(reliability)并不能直接等同，概念的混淆导致了研究者对个体差异测量可信度的忽视。

近年来，随着可重复性问题的提出(胡传鹏 等, 2016; Baker, 2016; Schlegelmilch, 2015)，一些研究者开始重新审视认知任务在测量个体差异时的测量学特征(如，Elliott et al., 2020; Enkavi et al., 2019; Hedge et al., 2018; Parsons et al., 2019)。研究发现尽管认知任务的群体信度较高，但其个体测量学信度却不如人意。例如，Hedge 等(2018)对七项经典认知任务的内类相关系数(ICC)进行系统评估，结果表明，这些任务有稳定的实验效应，但个体水平的信度很低(信度范围从 0 到 0.82 不等)。尽管这一现象近年来引起了广泛关注，但关于认知任务低信度的问题尚缺乏系统总结。同时，由于认知任务与传统问卷任务在结构上的差异，如何合理的对认知任务的信度进行评估也未得到充分的论述。

基于此，本文首先将概述信度概念，介绍适用于认知任务数据特点的信度计算方法；随后汇总现有认知任务信度研究，并分析信度分析中的潜在挑战；最后将探讨影响认知任务信度的因素及提升策略，以期为后续研究提供指导。

2 信度的定义及其影响因素

信度概念源于心理测量的经典测量理论(Classical Test Theory，CTT)。该理论将测量结果视为真分数和误差分数的组合，其中真实分数反映了被测量个体的真实特征，而误差分数则是由各种外部因素引起的随机波动(Lord et al., 1968; Xu et al., 2023)。信度衡量的是测验的一致性和稳定性，即测量工具在不同时间、不同环境下对同一特征的重复测量的一致程度(Crocker & Algina, 1986)。其公式通常表示为：

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2} = \frac{\sigma_X^2-\sigma_E^2}{\sigma_X^2}$$

其中，σ_T^2表示真分数的方差，即个体在该测量中的真实变异性；σ_X^2表示观测分数的方差，包括真分数的方差(σ_T^2)和测量之外因素导致的变化，即误差分数的方差(σ_E^2)。因此，信度反映了在观测分数中真实的个人差异的占比。信度越高, 意味着真分数变异的占比越高, 误差分数变异的占比越少，个体差异分析的有效性越高。反之，信度值低则表明测量工具无法有效地捕捉到个体差异。

在认知实验研究中，研究者通常通过分析被试的任务表现(包括选择结果、反应时间等行为指标)以及不同实验条件间的差异来探究目标效应。但测量个体差异时，信度评估面临着两个重要的挑战：一是认知任务的多变式和多条件性，且试次较多(如图1，以自我匹配任务(Sui et al., 2012)为例)，不同于传统问卷测量中每个被试在每个条目上仅有一个得分，要求信度评估方法与其适配。二是认知任务的实验效应往往可以从多个直接(如反应时、正确率)和衍生(如效率、认知模型参数)指标中进行选择，而不同指标的信度可能不尽相同。针对这些问题，下文将系统阐述认知任务信度的计算方法及如何选择其评估指标。

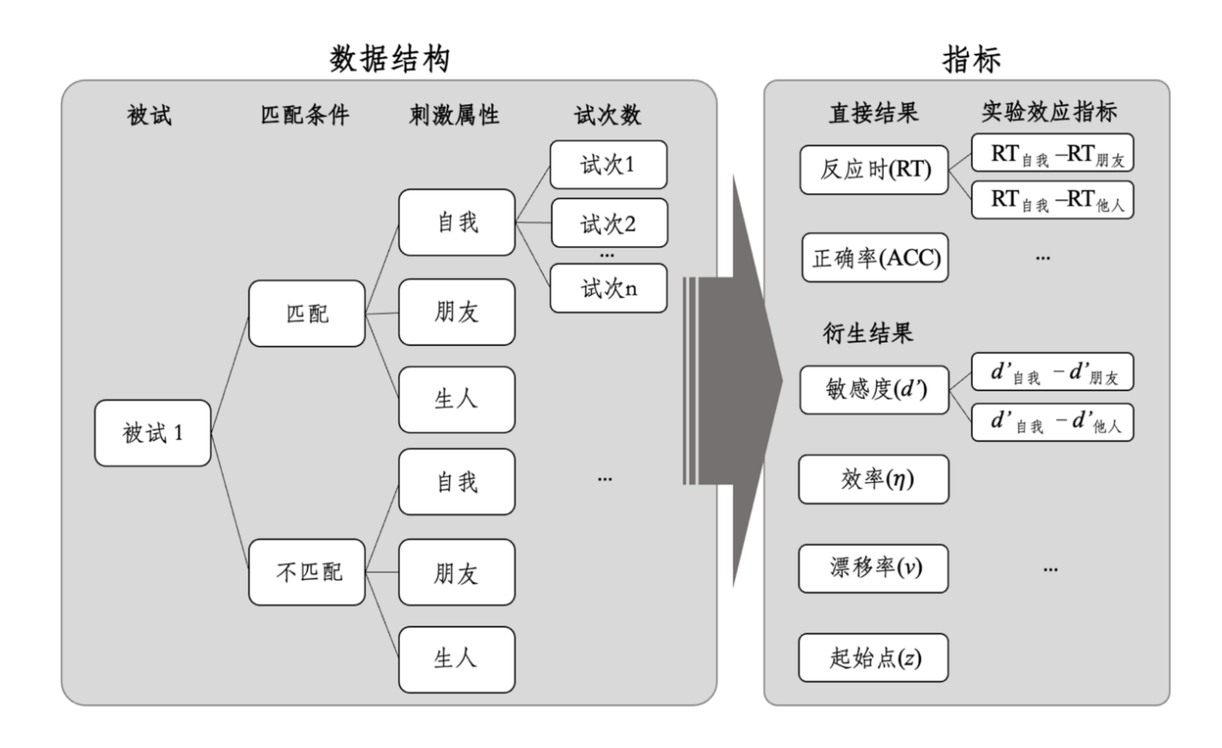


图 1.自我匹配任务的数据结构及指标. 敏感度指标基于信号检测理论的 *d*'；效率衡量反应时与正确率的权衡，计算公式为 RT / ACC；漂移率(*v*)和起始点(*z*)为漂移扩散模型的两个关键参数，分别表示证据积累的平均速率和积累起始点的位置(即是否更接近某个决策边界，详见刘逸康, 胡传鹏, 2023)。

3 认知任务信度的评估指标

评估认知任务信度面临的首要挑战是其数据结构的独特性。认知任务通常采用被试内设计，即每个被试在每个实验条件下均会完成多个试次，不同于传统问卷的每个条目单一得分的单点数据结构。因此，传统的内部一致性系数(Cronbach's α 系数)不适用于认知任务信度评估(Kahveci et al., 2024; Parsons et al., 2019)。具体而言，应用 Cronbach's α 计算信度需对试次数据简单平均，这将忽视试次间的变异，导致对内部一致性的评估不够准确(Kahveci et al., 2024)。此外，Cronbach's α 的计算依赖于所有条件对真实得分等贡献的假设，即 tau 等值预设(Novick & Lewis, 1967)。但这在认知任务中未必成立，可能导致较大的估计偏差(Parsons et al., 2019)。

据此，研究者建议采用两类信度评估指标(Kahveci et al., 2024; Pronk et al., 2022)：分半信度系数，以衡量个体在任务执行中不同试次间的一致性；重测信度，以评估个体在认知任务中时间点间的表现稳定性。

3.1 认知任务的分半信度的计算

在经典测量理论中，分半信度通过将测试等分为两半并计算得分相关性来评估内部一致性。认知任务的多条件和多试次特点要求研究者选择有效的分半方法，以获得可靠的估计。Pronk 等(2022)系统评估了四种常用的分半方法(见表 1)，发现置换分半法最为稳健。置换分半法通过多次无放回的随机抽样将实验试次分为两半，并计算每次分半后 Spearman-Brown 矫正的皮尔逊相关系数，得到信度系数分布。这种方法减少了单次分半可能带来的抽样误差，同时考虑到了实验设计中潜在的时间效应、任务设计效应以及非线性转换带来的影响，提供更为稳健的信度估计(Pronk et al., 2022)。后续研究发现，该方法能够广泛应用于多种认知任务(Ivanov et al., 2023; Liu et al., 2025; Zhang et al., 2023)。

目前，以上的多种分半计算信度的方法均能在 R 包 splithalfr 上实现(Pronk, 2020)。见在线补充材料中以自我匹配任务的反应时为例(Liu et al., 2025)，计算前后分半、奇偶分半以及置换分半的代码(https://www.scidb.cn/en/anonymous/QVJqbWFt)。

表 1 不同分半方法的优点、缺点及适用条件

分半方法	描述	混淆效应				优点	缺点	适用性
		时间效应	任务设计效应	试次抽样效应	非线性评分			
前后分半 (first–second methods)	根据试次序号分为前后两半	✗	✓	✗	✗	简单易实现	与时间效应(例如参与者的疲劳或学习效应)的混淆	当时间效应不显著或可控时可使用，但如果任务可能涉及学习或疲劳效应，应避免使用此方法
奇偶分半 (odd–even methods)	根据试次序号分为奇偶两半	✓	✗	✗	✗	有效控制时间效应	可能与任务设计(例如任务包含交替的条件)的混淆	在任务设计中如果存在交替条件时可能导致偏倚，需要谨慎使用
置换分半 (permutated methods)	无放回的随机分半	✓	✓	✓	✓	多次抽样，结果稳健	几乎没有	能较好应对解决试次抽样效应，适用于需要通过多次抽样来获得稳定估计的情况
蒙特卡洛分半 (Monte Carlo methods)	有放回的随机分半	✓	✓	✓	✓	不基于任务得分和试次之间存在线性关系的前提假设	计算量大，且当试次数少时，重复抽样导致结果方差降低，从而可能高估信度 (Kahveci et al., 2024)	适用于采用非线性评分的任务，但对于试次数较少的实验设计需谨慎

3.2 认知任务的重测信度的计算

重测信度(test-retest reliability)主要用于评估同一测量工具在不同时间点对同一被试重复测量的结果一致性。早期研究常用皮尔逊相关系数、配对样本 *t* 检验和 Bland-Altman 图等方法来评估重测信度(Bruton et al., 2000)。然而，这些方法均存在一定的局限性：皮尔逊相关系数仅反映线性相关性，配对 *t* 检验仅关注均值差异，而 Bland-Altman 图侧重一致性界限的可视化，因此这些方法均未能全面反映重测信度的核心特征(Koo & Li, 2016)。相比之下，内类相关系数(Intraclass Correlation Coefficient, ICC)能够同时反映测量结果的相关性和一致性，因而成为近年来评估重测信度的更优选择。

在实践中，存在多种 ICC 的计算方法。McGraw 和 Wong(1996)界定了 10 种 ICC 的变体，但这也使得据情境选择合适的变体变得困难。在认知任务的重测信度计算中，研究者指出最为推荐的是使用双向随机效应模型 ICC(2,1)和双向混合效应模型 ICC(3,1)(Parsons et al., 2019)。相关 ICC 的公式如下：

$$\mathrm{ICC2} = \frac{\mathrm{MS_{BS}} - \mathrm{MS_E}}{\mathrm{MS_{BS}} + (\mathrm{k}-1)\mathrm{MS_E} + \frac{\mathrm{k}}{\mathrm{n}}(\mathrm{MS_{BM}} - \mathrm{MS_E})} \qquad (1)$$

$$\mathrm{ICC3} = \frac{\mathrm{MS_{BS}} - \mathrm{MS_E}}{\mathrm{MS_{BS}} + (\mathrm{k}-1)\mathrm{MS_E}} \qquad (2)$$

其中，MS_{BS} 是被试间方差(Between-subject Mean Square)，反映个体间存在的差异；MS_E 是误差方差(Error Mean Square)，反映随机误差或噪声；MS_{BM} 是测量间方差，通常反映重测任务间的变异性；k 是试次；n 是被试总数。

Koo 和 Li(2016)提出，在选取重测信度的 ICC 模型时，应根据研究目标选用 ICC(2,1)或 ICC(3,1)。对于需将结果推广至更广泛群体的随机抽样研究，应采用 ICC(2,1)双向随机效应模型；若结果仅限于特定被试，则适用 ICC(3,1)双向混合效应模型。Parsons 等(2019)则建议应同时报告这两种信度，以进行对比。

在评估 ICC 时，需要注意的是，尚无统一样本量标准确定可靠性水平。但据通用的经验法则，研究者进行可靠性检验时，样本量需要至少达到 30(Koo & Li, 2016)。ICC 的解释通常遵循以下标准：ICC<0.5 表示可靠性差，0.5 – 0.75 为中等，0.75 – 0.9 为良好，>0.90 为优秀(Cicchetti & Sparrow, 1981; Kupper & Hafner, 1989)。然而，这些标准仅为一般性指导，具体的解释应结合研究的背景、测量工具和数据的特性。

目前，ICC(2,1)和 ICC(3,1)均能通过 R 包 psych 来实现(William Revelle, 2024)，见在线补充材料中以自我匹配任务的反应时为例(Liu et al., 2025)，计算 ICC 的代码。

4 估计认知任务信度的实验效应指标选取

评估认知任务信度的第二个挑战来自于实验效应指标的多样性。以经典的 Stroop 任务为例，其数据结果通常包括反应时(reaction times, RT)和错误率(Error rate)。由于 Stroop 任务通常包括一致条件(如“红”字用红色书写)和不一致条件(如“红”字用绿色书写)，此时，研究者要决定如何使用不同条件下的数据来估计信度。可选的数据指标至少有六种：一致/不一致条件的反应时、两种条件反应时之差、一致/不一致条件的错误率，以及两种条件错误率之差。此外，还存在潜在的衍生指标，例如将反应时间与错误率结合的效率指标(Efficiency)。尽管这些指标都可以用来衡量个体表现，但它们可能反映了不同的认知过程，因此用于估计信度时也会有不同的表现。

已有研究者发现，不同指标在信度估计中的表现可能存在显著差异。Hedge 等(2018) 对七种经典认知任务的研究显示，各任务内部指标的信度差异范围广泛(0-0.82)。例如，在 Stroop 任务中，反应时相关指标的重测信度通常高于错误率相关指标，一致条件下的反应时信度为 0.77，而错误率为 0.36。类似的，Liu 等(2025)在自我匹配任务中也发现，反应时和效率指标在内部一致性和重测信度上优于其他指标(如正确率、敏感度、扩散模型参数)。

然而，反应时间并不总是在所有认知任务中都显示出信度上的优势。von Bastian 等(2020)系统分析了注意控制领域不同类别（如抑制、转换、心智游离）的任务及其指标(反应时、准确性、敏感性等)的信度，未发现一致性规律。该研究发现，认知任务信度随任务类型显著变化。例如，心智游离(Mind Wandering)类认知任务下的的信度中位数较高(0.90；M = 0.89，SD = 0.08)，而抑制(Inhibition)类任务的信度中位数相对较低(0.79；M = 0.72，SD = 0.22)，且抑制类任务的信度系数分布范围更大。这表明抑制任务的信度可能受任务设计和执行特征影响，导致较大波动。

以上研究表明，在利用认知任务测量个体差异时，选择合适的实验效应指标至关重要。为便于研究者了解认知任务指标的信度表现，下表整理并呈现了一些对常见认知任务指标信度的综合比较的文章结果。

表 2.部分常见认知任务信度的综合比较文献及主要结论

源文献	主要任务	信度类型	主要结论
Elliott et al., 2020	常见 fMRI 任务 (如工作记忆、情绪加工、决策)	重测信度(ICC)	基于 90 个实验、1008 名被试的元分析结果显示，用于 fMRI 研究的认知任务的信度较低（ICC 均值 = 0.397）。此外，对 HCP（N = 45）和 Dunedin（N = 20）中的认知任务数据的分析进一步表明，11 种常见任务的 ICC 介于 0.067 至 0.485 之间，难以用于个体差异研究。
Enkavi et al., 2019	自我调节相关的行为任务 & 量表	重测信度(ICC)	根据 154 篇文章和 17550 人的数据的分析结果，行为任务的重测信度（0.610）平均低于自评量表的信度（0.716），且随着样本量的增加，信度呈下降趋势。基于 150 人的新数据显示，行为任务的重测信度平均低于自评量表（行为−自评的差异为−0.432，95% CI = [−0.482, −0.384]）。自评量表的中位 ICC 为 0.674，而行为任务的中位 ICC 为 0.311。
Hedge et al., 2018	7 种经典认知任务（Flanker、Stroop、stop-signal、go/no-go、Posner cueing、Navon、SNARC）	重测信度(ICC)	基于三个研究的分析结果显示，7 种经典认知任务的重测信度范围从 0 到 0.82，大多数任务的信度低于预期。
Huang et al., 2024	与冲动性相关的自评量表和行为任务(48 种冲动性测量，来源于 10 个自评量表和 10 个行为任务)	重测信度(ICC)	基于 1676 名被试的冲动性测量数据，采用双因素模型(bifactor model)对冲动性进行心理测量建模，提取出一个广义因子 I 和六个特定因子。CFA 结果表明，该模型拟合良好(SRMR = 0.06，RMSEA = 0.06，CFI = 0.93，TLI = 0.93)。通过对 198 名被试的间隔三个月的重测，CFA 结果也支持该模型(SRMR = 0.08，RMSEA = 0.04，CFI = 0.97，TLI = 0.97)。研究同时发现传统的自评量表的平均重测信度为 0.66，行为任务为 0.44。

Karvelis et al., 2023	计算精神病学中的多种任务（如赌博任务，联结学习任务）	Pearson 相关系数，重测信度 (ICC)	基于对 20 篇文章的文献综述，许多计算模型的参数的信度较低。文章提出了一系列改进建议，以确保计算测量方法能够成功应用于临床实践。

5 提高认知任务信度的途径

大量研究揭示了认知任务信度偏低的问题，使得信度成为认知科学和心理测量领域的共同议题(Zorowitz & Niv, 2023)。研究者通过分析表明，低信度主要归因于被试间变异性小于数据噪音。例如，Rouder 等(2023)重新分析了 Flanker 任务和 Stroop 任务的数据，发现试次间的噪音是被试间变异的 8 倍之多。因此，基于经典测量理论，优化信度可以从两方面入手：通过增加被试间变异性或降低测量误差，具体改进策略涵盖实验设计与计算统计方法。表 3 汇总了提升信度的策略及相关研究(Karvelis et al., 2023; Parsons et al., 2019; Zorowitz & Niv, 2023)，以供参考。

表 3 提高信度的主要策略及相关研究

主要策略分类	具体策略	策略描述	相关研究
增加被试间差异	调整任务难度	调整任务难度，避免天花板/地板效应，使得任务分数在不同被试间具有足够的变异性	Kyllonen et al., 2019; Oswald et al., 2015
	加入游戏化设计	通过融入角色扮演、互动反馈机制等元素，增加任务复杂性	Allen et al., 2024; Kucina et al., 2023; Sailer et al., 2017
	增加样本多样性	采用更广泛的被试来源，避免单一群体的认知特征限制	Arnon, 2020; Henrich et al., 2010; Kyllonen et al., 2019
减少测量误差	增加正式实验试次数量	通过增加试次减少随机误差，但需避免疲劳效应	McLean et al., 2018; Liu et al., 2025
	增加练习	通过充分练习帮助被试达到稳定的表现状态	Alexander et al., 2003; McLean et al., 2018
	控制实验环境	确保受试者注意力集中，减少外部干扰	Bruder et al., 2021
改良计算与统计方法	选取可靠的实验效应指标	选择更具信度的实验效应指标	Ross et al., 2015; Saville et al., 2011; Weigard et al., 2021
	计算建模	采用计算模型提取的参数提高信度，通过更精细的建模揭示个体差异	Sullivan-Toole et al., 2022; Xu & Stocco, 2021

使用分层模型	使用分层模型、贝叶斯分层模型、交叉混合效应模型等方法提高信度估计的准确性	潘晚坷 等, 2023; Haines et al., 2023; Rouder & Haaf, 2019
使用新信度指标	通过引入其他信度指标，克服传统信度分析的局限，以更全面和灵活的方式评估测量的稳定性、个体可区分性和任务本身的可测性	Rouder & Mehrvarz, 2024; Xu et al., 2023

5.1 增加被试间差异的策略

增加被试间的表现差异的核心在于更准确地捕捉个体差异，包括以下将几种策略。

5.1.1 调整任务难度

合理设置任务难度是确保认知任务信度的重要因素。任务过易或过难可能引起天花板或地板效应，掩盖个体能力差异(Oswald et al., 2015)。因此，设计认知任务时，研究者应确保任务难度与被试群体平均能力水平相匹配(Kyllonen et al., 2019)。在其能力水平未知的情况下，设计难度跨度较大的任务或筛选出引起极端分数的任务模块，是增加被试间变异性的有效方法(Feldt, 1993; Oswald et al., 2015)。

5.1.2 加入游戏化设计

提升被试参与度和投入感是增强认知任务信度的另一策略。游戏化通过融入游戏设计元素(如角色扮演、叙事背景、互动反馈机制以及协作与竞争等)，使传统任务呈现游戏特征(Sailer et al., 2017)。Kucina 等(2023)的研究表明，游戏化可以增加任务复杂性，激发多样化反应，从而提高信度。此外，整合游戏元素或视频刺激可提升参与度和动机，减轻疲劳效应(Allen et al., 2024)。然而，游戏化提升信度的有效性尚存争议。关键在于游戏化元素需有效吸引注意力以激励被试。若被试不接受叙事背景或跳过环节，就可能会降低统计功效，影响信度(Kucina et al., 2023; Sailer et al., 2017)。因此，在应用游戏化手段时，确保被试对游戏化元素的感知和参与是至关重要的。

5.1.3 增加样本多样性

样本同质性会减少被试间变异性。通过招募社区或在线平台样本(如 Amazon Mechanical Turk、Prolific Academic)等方式，提高被试群体的异质性(Arnon, 2020; Kyllonen et al., 2019)，可以增加被试间变异性。在增加样本多样性的同时，需要确保任务在不同群体间的测量等值性(measurement invariance)(Molenaar & Feskens, 2024)。例如，许多心理学研究基于 WEIRD(Western, Educated, Industrialized, Rich, Democratic)群体(刘伟彪 等, 2024; Ghai et al., 2024; Henrich et al., 2010)，认知任务是否会受到文化背景的影响可能是需要首先解决的问题，通过测量等值性检验（如多组结构方程模型或项目反应理论分析）确保跨文化信度的前提(Yarkoni, 2022)。

5.2 降低误差分数方差的策略

降低误差分数的方差的核心在于减少随机误差对测量结果的影响，包括如下几种策略。

5.2.1 增加正式实验的试次数量

增加试次数量是提高测量分数方差的有效策略。例如，在反应时任务中，增加试次可稳定被试的任务表现并扩大个体差异(McLean et al., 2018)。Liu 等(2025)在自我匹配任务中也发现，试次数与分半信度显著相关，试次越多，信度越高。此外，研究者还可以使用 Spearman-Brown 预测公式(Sanders et al., 1989)估算所需试次数。

5.2.2 增加练习

练习对任务信度评估至关重要。Alexander 等(2003)指出，在正式实验前充分练习可稳定被试表现，减少测量误差，提高信度；而缺乏充分练习可能导致被试表现前后差异较大，降低信度。McLean 等(2018)建议设置独立练习模块，并对练习数据进行独立分析或排除。

5.2.3 控制实验环境

实验环境一致性对测量信度至关重要。在实验室环境中，需标准化物理条件（如光照、噪音、座椅高度）；而在线测试中，可以通过注意力检查（如眼动追踪、反应时筛选）识别并排除低质量数据，增强数据可靠性(Pronk et al., 2023)。此外，虚拟现实（VR）技术通过提供沉浸式环境和严格实验控制，减少外界干扰，提高生态效度(Bruder et al., 2021)。值得注意的是，对于涉及不同设备(如智能手机与桌面端)的测量任务，需进行测量等值性检验，以确保测量信度和效度一致性(Pronk et al., 2023)。

5.3 改良计算与统计方法

研究者提出了多种计算与统计方法改良策略，包括选取可靠的实验效应指标、合适的认知建模计算模型以及采用层级模型等。这些方法旨在更精确地捕捉个体差异，增强研究结果的解释力和普遍性。

5.3.1 选取可靠的实验效应指标

在实验研究中，选择合适的效应指标对于确保结果的可靠性至关重要。在传统的实验研究中，认知任务通常通过计算个体在不同条件下的差异分数(如 Stroop 任务中"一致"与"不一致"条件下的反应时间差异)来评估效应。但差异分数的使用有其局限性。当条件测量指标高度相关且方差相似时，由于测量误差会在差异计算中累积，差异分数信度可能低于原始测量(Cronbach & Furby, 1970; Edwards, 2001; Lord, 1956)。此外，低信度的构成变量及其正相关关系会进一步降低差异分数信度(Edwards, 1994)。此外，差异分数的信度还受到构成变量信度的影响：如果构成变量的信度较低且二者之间存在正相关关系，差异分数的信度将进一步降低(Edwards, 1994)。

为规避差异分数的局限性，可采取以下策略：一是使用原始分数的简单平均(Ross et al., 2015)。二是探索替代效应指标，如个体内反应时间变异性(Intraindividual Response Time Variability)和认知效率(Cognitive Efficiency)，这些指标在执行控制研究中显示出更高信度(Saville et al., 2011; Weigard et al., 2021)。

5.3.2 计算建模

计算建模被广泛应用于认知任务研究，以更精细的模型参数揭示个体差异。研究表明，计算模型参数相较于传统行为指标，在特定情境下具有更高信度(Rappaport et al., 2025; Sullivan-Toole et al., 2022; Xu & Stocco, 2021)。例如，Rappaport 等(2025)在 Eriksen flanker 任务中，发现 DDM 参数（决策阈值、漂移率、非决策时间）比反应时和正确率等传统指标在不同样本间展现出更高的信度和一致性，且有效区分认知控制和决策差异。

然而，也有研究发现，计算模型参数信度并非总是高于直接任务结果，有时甚至被评为中等或更低(Hitchcock et al., 2022; Liu et al., 2025; Pike et al., 2022)。例如，Liu 等(2025) 在自我匹配任务中评估了标准扩散漂移模型 (DDM)在自我匹配任务 (SMT)中的适用性，发现漂移率 (*v*) 和起始点 (*z*) 的分半信度 和重测信度均低于可接受水平，表明其在衡量个体差异时缺乏稳定性。模型参数信度不足主要归因于模型复杂度导致的共线性问题和模型适用性。当模型包含过多的自由度和参数时，会提高共线性，导致过拟合，影响模型泛化能力，使其在解释新数据时表现不佳(Enkavi et al., 2019; Rey-Mermet et al., 2019)。此外，计算建模的有效性依赖于模型是否准确反映认知过程，不当的模型应用可能降低参数信度(Eckstein et al., 2022; 刘逸康, 胡传鹏, 2023)。因此，研究者需要结合理论背景和任务特性，谨慎选择或开发适合的模型，以确保模型参数能够有效反映个体差异。

5.3.3 使用分层模型

针对认知任务数据的分层特性，研究者建议使用分层模型(hierarchical models)以区分试次间和个体间变异，减少试次变异对个体差异估计的干扰(Rouder & Haaf, 2019)。此外，分层模型因其独立于试次数量，还可能提高结果的可推广性，增强不同实验设计结果的可比性(Rouder & Haaf, 2019)。方法上，贝叶斯分层模型(Bayesian hierarchical models)用于信度估计时，可以避免经典统计中参数估计的问题(Haines et al., 2023; 潘晚坷等., 2023)。当数据结构更复杂时，研究者可以考虑使用交叉混合效应模型(Crossed Random Effects Model)和混合效应位置尺度模型(Mixed-Effects Location Scale Model, MELSM)(Brunton-Smith et al., 2017; Williams et al., 2021)。尽管分层模型可能能

够更好地估计信度，其也存在局限。例如，贝叶斯分层模型的估计结果可能对先验设置敏感，需要研究者谨慎选择先验分布和参数(Katahira et al., 2024)。此外，由于分层模型的参数普遍较多，需要较多试次数，增加了实现的难度。

5.3.4 使用新信度指标

Xu 等(2023)等提出了新的信度指标，包括非参数信度指标 (nonparametric reliability indices)和多变量信度计算方法(multivariate reliability estimation methods)，扩展了传统信度分析。非参数信度指标中，个体区分度(discriminability)可以衡量测量的可区分性，即个体在不同测量间的变异是否大于其在不同测量中的变异；而指纹法(fingerprinting)则用于评估测量在不同时间点的稳定性，即同一被试在多次测量中的表现是否高度一致，其核心思想是以个体组内距离(within-individual distance)与组间距离(between-individual distance)的比值确定测量识别相同个体的稳定性。

在多变量信度计算方法方面，Xu 等(2023)提出了 ICC 的两个变式：基于距离的 ICC(distance-based intraclass correlation coefficient, dbICC)和基于信息论的信度(information-based intraclass correlation coefficient, I2C2)。dbICC 是基于欧几里得距离(Euclidean distance)计算的信度指标，通过比较个体在不同测量中的均值内距离(mean within-individual distance, MSD_x)与均值间距离(mean between-individual distance, MSD_v)来估算信度值，适用于多维度数据分析。I2C2 则利用协方差矩阵(covariance matrix)计算测量的稳定性，通过评估测量变量的总体变异与个体变异的比值。这些新指标超越了单变量信度测量的限制，提供了精细的个体识别和稳定性评估工具。Xu 等(2023)还开发了用于计算这些指标的 R 包 Rex，见在线补充材料中以自我匹配任务的反应时为例(Liu et al., 2025)的计算代码。

Rouder 和 Mehrvarz (2024)提出信号噪声比(Signal-to-Noise Ratio, SNR)作为评估任务可靠性的新指标，以减少对实验设计的依赖。SNR 通过层级模型比较个体间变异与试次内噪声，提供独立于试次数的可靠性评估。与传统信度系数相比，SNR 更直接反映任务可测性，允许独立于实验设计进行评估。SNR 有助于设定试次数量，确保测量精度，并判断任务间低相关性是否由可靠性限制导致。

6 总结与展望

量化人类认知过程是科学心理学的重要目标，并促进实验与相关研究取向的融合。同时，个体在认知过程上的异常可能是潜在干预精神疾病研究的靶点，受到精神病学的关注。美国国家精神卫生研究所(NIMH)提出的精神疾病研究领域标准(RDoC，Research Domain Criteria)框架中(Cuthbert, 2022; Cuthbert & Insel, 2013; Insel et al., 2010)，认知过程被视为核心领域之一。然而，若不解决认知任务的低信度问题，认知任务难以成为有效衡量个体差异的工具。

对认知任务信度的研究推动了相关领域的进展。现有研究提出了适应实验设计的传统信度估计方法(Koo & Li, 2016; Parsons et al., 2019; Pronk et al., 2022)和新的信度指标 (Rouder & Mehrvarz, 2024; T. Xu et al., 2023)，并利用其对经典认知任务的信度进行了评估。在此基础上，一些提高信度的潜在途径，如游戏化，已获得支持性证据，初步显示其有效性(Kucina et al., 2023)。

然而，如何让认知任务成为好的个体差异测量工具，仍然有许多亟待探索的研究问题。未来的研究可从以下几个方面继续探索：(1)对更广泛的认知任务进行信度评估(如 Enkavi et al., 2019; Karvelis et al., 2023)，而非仅局限于经典的认知任务；(2)改进现有认知任务(Liu et al., 2025; McLean et al., 2018)；(3)超越经典测量理论框架(Haines et al., 2023; Rouder & Mehrvarz, 2024)；(4)开发新的、适合测量个体差异的认知任务(Kucina et al., 2023)。这些探索将有望推动认知过程的个体差异测量的进一步发展，进而帮助更好地理解人类认知过程的内在机制与应用。

参考文献

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题：从危机到契机. *心理科学进展*, *24*(9), 1504.

刘伟彪, 陈志毅, 胡传鹏. (2024). 心理与脑科学研究中的样本代表性. *科学通报*, *69*(24), 3515–3531.

刘逸康, 胡传鹏. (2023). 证据积累模型的行为与认知神经证据. *科学通报*, *69*(8), 1068–1081.

潘晚坷, 温秀娟, 金海洋. (2023). 贝叶斯混合效应模型: 基于 brms 的应用教程. *心理技术与应用*, *11*(10), 577–598.

区健新, 吴寅, 刘金婷, 李红. (2020). 计算精神病学: 抑郁症研究和临床应用的新视角. *心理科学进展*, *28*(1), 111–127.

Alexander, C., Paul, M., & Michael, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*, *9*(3), 419–428.

Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., … Ho, M. K. (2024). Using games to understand the mind. *Nature Human Behaviour*, *8*(6), 1035–1043.

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, *52*(1), 68–81.

Baker, M. (2016). *1,500 scientists lift the lid on reproducibility*. Nature Publishing Group UK London.

Bruder, L. R., Scharer, L., & Peters, J. (2021). Reliability assessment of temporal discounting measures in virtual reality environments. *Scientific Reports*, *11*(1), 7015.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(2), 551–568.

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: what is it, and how is it measured? *Physiotherapy*, *86*(2), 94–99.

Chen, Z., & Chuan-Peng, H. (2024). *Sample Representativeness in Psychological and Brain Science Research*.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin*, *74*(1), 68.

Cuthbert, B. N. (2022). Research Domain Criteria (RDoC): Progress and Potential. *Current Directions in Psychological Science*, *31*(2), 107–114.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, *11*(1), 126.

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269.

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *Elife*, *11*, e75474.

Edwards, J. R. (1994). Regression Analysis as an Alternative to Difference Scores. *Journal of Management*, *20*(3), 683–689.

Edwards, J. R. (2001). Ten Difference Score Myths. *Organizational Research Methods*, *4*(3), 265–287.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., … Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477.

Feldt, L. S. (1993). The Relationship Between the Distribution of Item Difficulties and Test Reliability. *Applied Measurement in Education*, *6*(1), 37–48.

Geng, S., Liu, S., Fu, Z., Ge, Y., & Zhang, Y. (2022). Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems* (pp. 299–315). ACM.

Ghai, S., Forscher, P. S., & Chuan-Peng, H. (2024). Big-team science does not guarantee generalizability. *Nature Human Behaviour*, *8*(6), 1053–1056.

Haines, N., Sullivan-Toole, H., & Olino, T. (2023). From classical methods to generative models: Tackling the unreliability of neuroscientific measures in mental health research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 822–831.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational Psychiatry Needs Time and Context. *Annual Review of Psychology*, *73*(1), 243–270.

Huang, Y., Luan, S., Wu, B., Li, Y., Wu, J., Chen, W., & Hertwig, R. (2024). Impulsivity is a stable, measurable, and predictive psychological trait. *Proceedings of the National Academy of Sciences*, *121*(24), e2321758121.

Huys, Q. J., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, *46*(1), 3–19.

Huys Q.J. M. (2015). Computational psychiatry. In Jaeger D. & Jung R. (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 775–783). Springer.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., … Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751.

Ivanov, Y., Theeuwes, J., & Bogaerts, L. (2023). Reliability of individual differences in distractor suppression driven by statistical learning. *Behavior Research Methods*, *56*(3), 2437–2451.

Kahveci, S., Bathke, A. C., & Blechert, J. (2024). Reaction-time task reliability is more accurately computed with permutation-based split-half correlations than with Cronbach's alpha. *Psychonomic Bulletin & Review*. Advance online publication.

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*, *148*, 105137.

Katahira, K., Oba, T., & Toyama, A. (2024). Does the reliability of computational models truly improve with hierarchical modeling? Some recommendations and considerations for the assessment of model parameter reliability: Reliability of computational model parameters. *Psychonomic Bulletin & Review*, *31*(6), 2465–2486.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.

Kucina, T., Wells, L., Lewis, I., de Salas, K., Kohl, A., Palmer, M. A., … Heathcote, A. (2023). Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. *Nature Communications*, *14*(1), 2234.

Kupper, L. L., & Hafner, K. B. (1989). On assessing interrater agreement for multiple attribute responses. *Biometrics*, 957–967.

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., … Stankov, L. (2019). General fluid/inductive reasoning battery for a high-ability population. *Behavior Research Methods*, *51*(2), 507–522.

Liu, Z., Hu, M., Zheng, Y., Sui, J., & Chuan-Peng, H. (2025). A multiverse assessment of the reliability of the self-matching task as a measurement of the self-prioritization effect. *Behavior Research Methods*, *57*(1), 37.

Lord, F. M. (1956). Sampling Error due to Choice of Split in Split-Half Reliability Coefficients. *The Journal of Experimental Education*, *24*(3), 245–249.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.*

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30.

McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2018). Towards a reliable repeated-measures beads task for assessing the jumping to conclusions bias. *Psychiatry Research*, *265*, 200–207.

Molenaar, D., & Feskens, R. (2024). Relating violations of measurement invariance to group differences in response times. *Psychological Methods*.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*(1), 1–13.

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343–1355.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395.

Pike, A. C., Tan, K., Ansari, H. J., Wing, M., & Robinson, O. J. (2022). *Test-retest reliability of affective bias tasks*.

Pronk, Hirst, R. J., & Wiers, R. W. (2023). Can we measure individual differences in cognitive measures reliably via smartphones? A comparison of the flanker effect across device types and samples. *Behavior Research Methods*, *55*(4), 1641–1652.

Pronk, T. (2020). Splithalfr: Extensible bootstrapped split-half reliabilities. *R Package Version*, *2*, 12.

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44–54.

Rappaport, B. I., Shankman, S. A., Glazer, J. E., Buchanan, S. N., Weinberg, A., & Letkiewicz, A. M. (2025). Psychometrics of drift-diffusion model parameters derived from the Eriksen flanker task: Reliability and validity in two independent samples. *Cognitive, Affective, & Behavioral Neuroscience*, *25*(2), 311–328.

Rey-Mermet, A., Gade, M., Souza, A. S., Von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General*, *148*(8), 1335.

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods*, *47*(3), 736–743.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467.

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*, *30*(6), 2049–2066.

Rouder, J. N., & Mehrvarz, M. (2024). Hierarchical-Model Insights for Planning and Interpreting Individual-Difference Studies of Cognitive Abilities. *Current Directions in Psychological Science*, *33*(2), 128–135.

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380.

Sanders, P. F., Theunissen, T., & Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika*, *54*(4), 587–598.

Saville, C. W., Pawling, R., Trullinger, M., Daley, D., Intriligator, J., & Klein, C. (2011). On the stability of instability: Optimising the reliability of intra-subject variability of reaction times. *Personality and Individual Differences*, *51*(2), 148–153.

Schlegelmilch, R. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716–aac4716.

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1105.

Sullivan-Toole, H., Haines, N., Dale, K., & Olino, T. M. (2022). Enhancing the psychometric properties of the iowa gambling task using full generative modeling. *Computational Psychiatry*, *6*(1), 189.

Vasey, M. W., Dalgleish, T., & Silverman, W. K. (2003). Research on Information-Processing Factors in Child and Adolescent Psychopathology: A Critical Commentary. *Journal of Clinical Child & Adolescent Psychology*, *32*(1), 81–93.

von Bastian, C. C., Blais, C., Brewer, G., Gyurkovics, M., Hedge, C., Kałamała, P., … Rouder, J. N. (2020). *Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations*.

Weigard, A., Clark, D. A., & Sripada, C. (2021). Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition*, *215*, 104818.

William Revelle. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research* (p. 2.4.12). Northwestern University.

Williams, D. R., Martin, S. R., & Rast, P. (2021). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behavior Research Methods*, *54*(3), 1272–1290.

Xu, T., Kiar, G., Cho, J. W., Bridgeford, E. W., Nikolaidis, A., Vogelstein, J. T., & Milham, M. P. (2023). ReX: an integrative tool for quantifying and optimizing measurement reliability for the study of individual differences. *Nature Methods*, *20*(7), 1025–1028.

Xu, Y., & Stocco, A. (2021). Recovering Reliable Idiographic Biological Parameters from Noisy Behavioral Data: the Case of Basal Ganglia Indices in the Probabilistic Selection Task. *Computational Brain & Behavior*, *4*(3), 318–334.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.

Yarkoni, T., & Braver, T. S. (2010). Cognitive Neuroscience Approaches to Individual Differences in Working Memory and Executive Control: Conceptual and Methodological Issues. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of Individual Differences in Cognition* (pp. 87–107). Springer New York.

Zhang, Z., Yang, L.-Z., Vékony, T., Wang, C., & Li, H. (2023). Split-half reliability estimates of an online card sorting task in a community sample of young and elderly adults. *Behavior Research Methods*, *56*(2), 1039–1051.

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 789–797.